1628–1635 *Nucleic Acids Research, 1998, Vol. 26, No. 7*

Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences

Timothy M. Rose, Emily R. Schultz, Jorja G. Henikoff[1], Shmuel Pietrokovski[1], Claire M. McCallum[1] and Steven Henikoff[1,2,*]

Department of Pathobiology, School of Public Health and Community Medicine, University of Washington, Seattle, WA 98195, USA and [1]Fred Hutchinson Cancer Research Center and [2]Howard Hughes Medical Institute, 1100 Fairview Avenue N, Seattle, WA 98109-1024, USA

Received December 22, 1997; Revised and Accepted February 12, 1998

ABSTRACT

We describe a new primer design strategy for PCR amplification of unknown targets that are related to multiply-aligned protein sequences. Each primer consists of a short 3′ degenerate core region and a longer 5′ consensus clamp region. Only 3–4 highly conserved amino acid residues are necessary for design of the core, which is stabilized by the clamp during annealing to template molecules. During later rounds of amplification, the non-degenerate clamp permits stable annealing to product molecules. We demonstrate the practical utility of this hybrid primer method by detection of diverse reverse transcriptase-like genes in a human genome, and by detection of C^5 DNA methyltransferase homologs in various plant DNAs. In each case, amplified products were sufficiently pure to be cloned without gel fractionation. This COnsensus-DEgenerate Hybrid Oligonucleotide Primer (CODEHOP) strategy has been implemented as a computer program that is accessible over the World Wide Web (http://blocks.fhcrc.org/codehop.html) and is directly linked from the BlockMaker multiple sequence alignment site for hybrid primer prediction beginning with a set of related protein sequences.

INTRODUCTION

Most applications of the polymerase chain reaction (PCR) are based on designing primers that precisely match a known target sequence. However, in some situations, primers are targeted to unknown sequences, as when trying to isolate genes encoding proteins that belong to known protein families (1–5). In such cases, PCR primer design is usually based on reverse translation of multiply aligned sequences across the conserved regions of proteins (blocks). Various rules of thumb have been applied to this problem, but frequent failures to amplify a desired target sequence are often attributable to inadequate primer design. Primer design can be very difficult because of codon degeneracy and the additional degeneracy needed to represent multiple codons at a position in the alignment. These degeneracies lead to complications in trying to find suitable annealing temperatures and primer lengths. The need to target regions of high sequence conservation containing codons of low degeneracy limits PCR detection of unknown sequences to fairly close relatives, so improvements in primer design have the potential to be widely applicable.

To isolate distantly-related sequences by PCR, two strategies have previously been employed. One is to synthesize a pool of degenerate primers containing most or all of the possible nucleotide sequences implicit in a multiple alignment (Fig. 1A). One problem with this approach is that as the degeneracy increases to accomodate more divergent genes, the concentration of any single primer drops. As a result, the number of primer molecules in a PCR reaction that can prime synthesis during the amplification cycles drops, and these primers are used up early in the reaction. In addition, artifactual amplification occurs because of the dominance of primers in the pool which do not participate in amplification of the targeted gene but are available to prime non-specific synthesis. These problems are exacerbated by the low stringency annealing conditions that may be needed to detect mismatched homologs, especially when using short primers required for short conserved blocks. The result is a weak or undetectable band on a gel that might be no higher than background. The second strategy is to design a single consensus primer across the highly conserved region. The consensus primer is usually derived by choosing the most common nucleotide at every position of multiply aligned nucleotide sequences. Although this technique has been most successful in the isolation of highly conserved gene homologs, primer-to-template mismatches preclude its application to distantly related sequences.

Here we describe a strategy that overcomes problems of both degenerate and consensus methods for primer design: COnsensus-DEgenerate Hybrid Oligonucleotide Primers (CODEHOP, Fig. 1B). Hybrid primers consist of a relatively short 3′ degenerate core and a 5′ non-degenerate consensus clamp. Reducing the length of the 3′ core to a minimum decreases the total number of individual primers in the degenerate primer pool. Hybridization of the 3′ degenerate core with the target template is stabilized by the 5′ non-degenerate consensus clamp, which allows higher annealing temperatures without increasing the degeneracy of the pool. Although potential mismatches may occur between the 5′ consensus clamp of the primer and the target sequence during the initial PCR cycles, they are situated away from the 3′ hydroxyl extension site, and so mismatches between the primer and the target are less disruptive to priming of polymerase extension. Further amplification of primed PCR products during subsequent rounds of primer hybridization and extension is enhanced by the

*To whom correspondence should be addressed. Tel: +1 206 667 4515; Fax: +1 206 667 5889; Email: steveh@muller.fhcrc.org

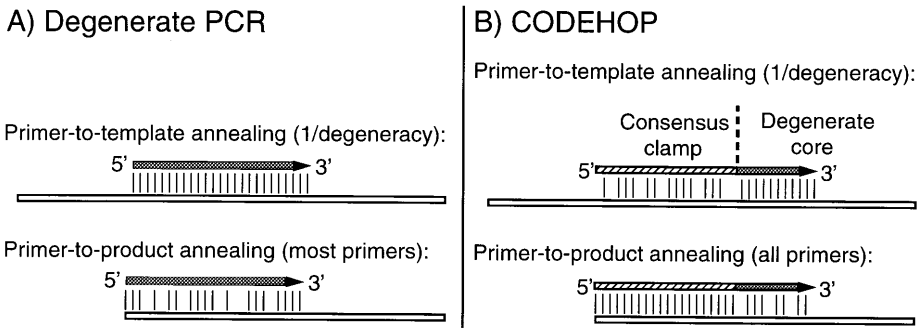

Figure 1. Schematic comparison of standard degenerate PCR (**A**) with the CODEHOP strategy (**B**), illustrating regions of mismatch in primer-to-template annealing during early PCR cycles and in primer-to-product annealing during subsequent cycles. Vertical lines indicate nucleotide matches between primer (arrow) and template or synthesized product. The overall degeneracy is the product of degeneracies at each nucleotide position, so that the fraction of precisely hybridizing primers = 1/degeneracy.

sequence similarity of all primers in the pool; this potentially allows utilization of all primers in the reaction.

We demonstrate the CODEHOP strategy by successfully amplifying unknown sequences from a background of genomic DNA. We also describe a program, implemented for the World Wide Web (WWW), for automatically predicting optimal primers that embody the CODEHOP strategy. The practical utility of this program is demonstrated by isolating members of a rapidly evolving family of novel cytosine methyltransferase homologs from diverse plants.

MATERIALS AND METHODS

Primer design

A CODEHOP is degenerate at the 3′ core region of length 11–12 bp across four codons of highly conserved amino acids and is non-degenerate at the 5′ consensus clamp region of 18–25 bp. Initially, such primers were designed by visual examination of protein multiple alignments made using ClustalW (6). This manual approach employing heuristic rules to identify suitable regions was later superseded by the development of a program that performs an exhaustive search. The CODEHOP program designs a pool of primers containing all possible 11- or 12mers for the 3′ degenerate core region and having the most probable nucleotide predicted for each position in the 5′ non-degenerate clamp region.

The program consists of the following steps. (i) A set of blocks is input, where a block is an aligned array of amino acid sequence segments without gaps that represents a highly conserved region of homologous proteins (7). A weight is provided for each sequence segment (8), which can be increased to favor the contribution of selected sequences in designing the primer. A codon usage table is chosen for the target genome. (ii) A position-specific scoring matrix is computed for each block using the odds-ratio method (9). (iii) A consensus amino acid residue is selected for each position of the block as the highest scoring amino acid in the matrix. (iv) For each position of the block, the most common codon corresponding to the amino acid chosen in step iii is selected utilizing the user-selected codon usage table (10). This selection is used for the default 5′ consensus clamp in step viii. (v) A DNA position-specific scoring matrix is calculated from the amino acid matrix (step ii) and the codon usage table. The DNA matrix has three positions for each position of the amino acid matrix. The score for each amino acid is divided among its codons in proportion to their relative weights from the codon usage table, and the scores for each of the four different nucleotides are combined in each DNA matrix position. Nucleotide positions are treated independently when the scores are combined. As an option, the highest scoring nucleotide residue from each position can replace the most common codons from step iv that are used in the consensus clamp. (vi) The degeneracy is determined at each position of the DNA matrix based on the number of bases found there. As an option, a weight threshold can be specified such that bases that contribute less than a minumum weight are ignored in determining degeneracy. (vii) Possible degenerate core regions are identified by scanning the DNA matrix in the 3′ to 5′ direction. A core region must start on an invariant 3′ nucleotide position, have a length of 11 or 12 positions ending on a codon boundary, and have a maximum degeneracy of 128 (current default). The degeneracy of a region is the product of the number of possible bases in each position. (viii) Candidate degenerate core regions are extended by addition of a 5′ consensus clamp from step iv or v. The length of the clamp is controlled by a melting point temperature calculation (11,12) (current default = 60 °C) and is usually ~20 nucleotides. (ix) Steps vii and viii are repeated on the reverse complement of the DNA matrix from step v for primers corresponding to the opposite DNA strand.

Molecular and sequence analysis

Primers were synthesized either commercially (Oligos Etc) or by the Hutchinson Center Biotechnology facility. Nucleic acids were extracted from macaque and human tissues and cell lines as described (13) and from *Arabidopsis* leaves using a Qiagen plant DNA kit. A set of crude plant DNAs was a gift from Amy Denton. Each 50 µl amplification reaction was performed using 25 pmol of each primer pool in a thin-walled 0.5 ml microcentrifuge tube in either a Perkin-Elmer 480 or MJ Research PTC100 thermal cycler. Whole PCR products were cloned using the TOPO-TA cloning kit (Invitrogen). Agarose gel analysis and DNA sequencing were performed using standard methods (14,15). Dendrograms were produced using the neighbor-joining and bootstrapping procedures in ClustalW (6) as implemented on the Blocks WWW site (16).

RESULTS

The hybrid primer strategy was tested on problems in which the target sequence for amplification was unknown but could be predicted from multiply aligned protein sequences. In the first

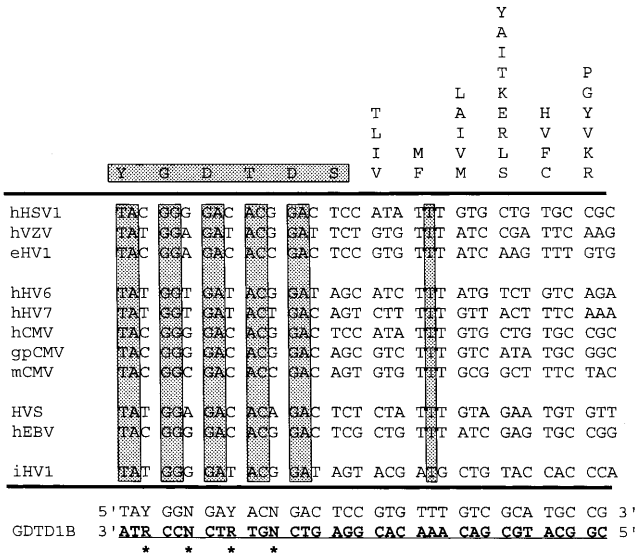

Figure 2. Hybrid primer design strategy for DNA polymerase genes of different herpes viruses. The nucleotide sequences across the conserved YGDTD sequence block from a variety of herpes virus DNA polymerase genes are aligned by codons. The invariant nucleotide positions are shown in shaded boxes. The amino acid sequences encoded at the various positions are shown on top with the YGDTD motif highlighted. The sequences are grouped within the α, β and γ subclasses of herpesviruses in descending order in the figure with the catfish herpes virus as an outlier. The GDTD1B hybrid primer pool was designed as a negative strand primer and is shown underlined. The IUBPAC codes for nucleotide degeneracies are used, and the degenerate positions are indicated (*). The primer pool is 64-fold degenerate, and each primer is 35 bp in length. (hHSV1, human herpes simplex virus 1 GenBank #X14112; hVZV, human varicella virus GenBank #X04370; eHV1, equine herpes virus 1 GenBank #M86664; hHV6, human herpes virus 6 GenBank #M63804; hHV7, human herpes virus 7 GenBank #U43400; hCMV, human cytomegalovirus GenBank #X17403; gpCMV, guinea pig cytomegalovirus, GenBank #L25706; mCMV, mouse cytomegalovirus GenBank #M73549; HVS, herpes virus saimiri #X64346; hEBV, human Epstein–Barr virus GenBank #V015555; iHV, ictalurid (catfish) herpes virus GenBank #M75136).

Figure 3. Hybrid primer design strategy for reverse transcriptase genes from various retroviral sequences. The nucleotide sequences across the conserved LQPG sequence blocks from a variety of retroviral sequences are aligned by codons. The invariant nucleotide positions are shown in open boxes. The amino acid sequences encoded at the various positions are shown on top with the evident LQPG motif highlighted. The sequences are grouped depending on the presence of a 'W', 'M' or 'F' codon immediately following the LQPG block, and the conserved nucleotides within these codons are shown in shaded boxes. The three hybrid primers designed from the 'W', 'M' and 'F' sequence groups are listed below with the degenerate positons indicated (*). (HIV1, human immunodeficiency virus type 1 GenBank #M38432; HIV2, human immunodeficiency virus type 2 GenBank #A05350; SIVAGM, simian immunodeficiency virus strain AGM GenBank #X07805; CAEV, caprine arthritis encephalitis virus GenBank #M33677; OMVV, ovine lentivirus GenBank #M31646; BIV, bovine immunodeficiency virus GenBank#M32690; FIV, feline immunodeficiency virus GenBank #M25381; SMRV, simian sarcoma virus GenBank #M23385; MMTV mouse mammary tumor virus #M15122; RSV, Rous Sarcoma Virus GenBank #J02342; HTLV1, human T-cell lymphotrophic virus 1 GenBank #L36905; HTLV2, human T-cell lymphotrophic virus 2 GenBank #L11456; HERSEQA, human endogenous retrovirus sequence GenBank #M96062; EIA, equine infectious anemia virus GenBank #U01866).

test, hybrid primers aimed at identifying a new primate herpes virus were designed from multiple sequence alignments of DNA polymerases from different herpes viruses. The second test used hybrid primers designed from alignments of reverse transcriptases from different retroviral genomes to identify a family of related retroviral elements within the human genome. In these tests, the hybrid primers were manually designed from multiple sequence alignments. The third test utilized the automated CODEHOP prediction program to design optimal primers from BlockMaker-generated alignments (17) of several DNA methyltransferases. Predicted CODEHOPs were used to identify members of a new subfamily of DNA methyltransferases from different plant genomes.

Detection of novel genomes using hybrid primers

We predicted that macaque retroperitoneal fibromatosis, a tumor similar to Kaposi's sarcoma, might contain a herpes virus homologous to the newly identified Kaposi's sarcoma-associated herpes virus (13). To identify and characterize such an unknown herpes virus, the amino acid sequences of the DNA polymerase genes (~1000 aa) from 11 different herpes virus genomes from the α, β and γ subclasses were multiply aligned. Visual examination of the alignment revealed five blocks that contained invariant regions suitable for primer prediction. Three blocks were chosen for primer design after evaluation of codon degeneracy within the blocks and distance between blocks. Primers were designed from these three regions using all codon possibilities for the 3′ degenerate core and the most frequent nucleotide in each position for the 5′ consensus clamp. The design strategy is shown for the most conserved sequence block (Fig. 2). As previously described (13), a hemi-nested PCR strategy was developed to use these three primers in two successive amplification reactions at 60°C to detect low amounts of viral DNA in a background of cellular genomic DNA from formalin-fixed paraffin-embedded samples. A PCR product of the correct size was detected on an electrophoretic gel. This product was cloned and sequenced and was shown to correspond to a DNA polymerase gene of a new macaque herpes virus most closely related to the human Kaposi's sarcoma-associated herpes virus (13). The success of the hybrid primer strategy in this example encouraged its refinement and extension to isolate other distantly-related sequences.

Isolation of homologous sequences from a multi-gene family within one genome using hybrid primers

To determine the nature and extent of retroviral sequence elements within the human genome, we designed primers to

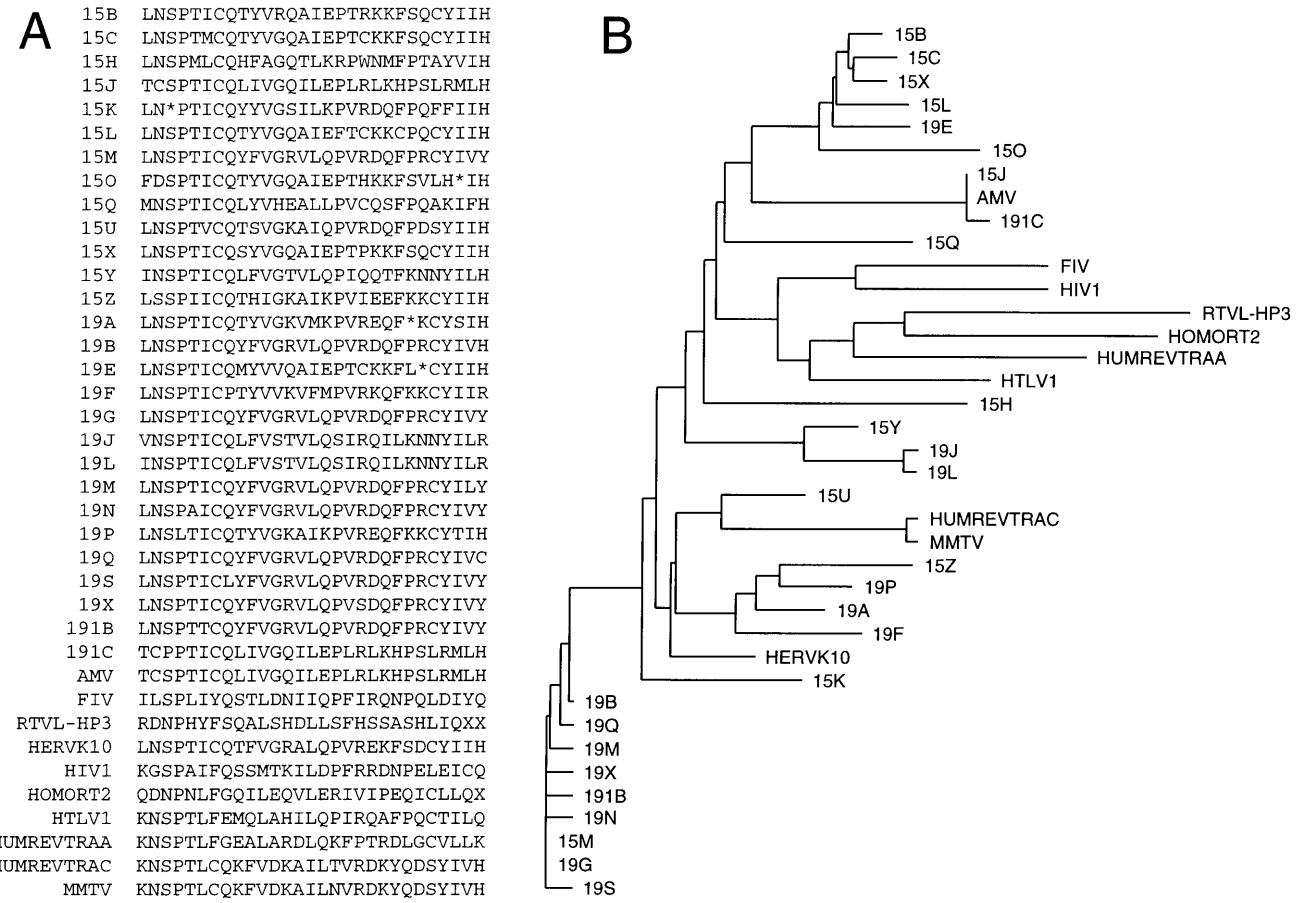

Figure 4. Alignment (**A**) and dendrogram (**B**) of amino acid sequences encoded by multiple endogenous reverse transcriptase-related sequences detected with hybrid primers LPQGM and YMDD from human tissue. Nucleic acids were prepared from paraffin blocks of lesions from Kaposi's sarcoma (clones designated 19) and rheumatoid arthritis (clones designated 15) using xylene washes and proteinase-K digestion as described (13). cDNA was synthesized using AMV reverse transcriptase with the hybrid primer pool (YMDD) predicted from the downstream YMDD motif. Amplification was performed using either of the upstream LPQGM, LPQGW or LPQGF hybrid primers (50 pmol) in combination with the downstream YMDD hybrid primer pool (50 pmol) in 0.067 M Tris buffer (pH 8.8), 4 mM $MgCl_2$, 16 mM $(NH_4)_2SO_4$, 10 mM 2-mercaptoethanol containing 100 µg bovine serum albumin per ml (16) for 35 cycles (1 min at 94°C, 1 min at 55°C, 1 min at 72°C). A hot start was obtained by initially incubating at 65°C prior to addition of Taq polymerase (2.5 U/50 µl). The amplification products were visualized on a 2.5% agarose gel with ethidium bromide and UV irradiation. The encoded amino acid sequences of series 19 and 15 cloned inserts (GenBank #AF047584–AF047597 and #AF050504–AF050516) are aligned with the corresponding sequences from 10 endogenous and viral reverse transcriptase sequences (RTVL-Hp3, GenPept #423062; HOMORT2 #257757; HERVK10, GenBank #M14123; HUMREVTRAA, #M25766; HUMREVTRAC, #M25768; AMV #S74099). Positions containing insertions or deletions in pseudogenes are indicated (*).

detect unknown reverse transcriptase-like sequences. The amino acid sequences of reverse transcriptase genes from 14 different retroviruses and retroviral sequences were multiply aligned. Two invariant sequence motifs (LPQG) and (YMDD) separated by ~40 aa (120 bp) were identified. The LPQG motif could be separated into three different sequence groups based on the identity of the amino acid immediately following the LPQG motif (M, W or F), and so three different hybrid primers were designed. The primer pools were 32-fold degenerate and 29–30 bp in length (Fig. 3). Amplification was performed at 55°C using the different combinations of the upstream hybrid primer pools (LPQGM, LPQGF, and LPQGW) and the downstream primer pool (YMDD), which was 24-fold degenerate and 30 bp in length.

Electrophoretic analysis revealed a single band of the expected size in the amplification reactions from the two tissue samples examined using the LPQGM and YMDD primers. No bands were detected using the LPQGF or LPQGW primers. The LPQGM-YMDD reaction mixtures were used for cloning, and 52 individual clones were sequenced, 26 from each of the two tissue sources. Forty-eight of the clones contained amplified products corresponding to reverse transcriptase coding regions, which are closely related to the mouse mammary tumor virus sequences. Twenty-seven different sequences were identified: four of these are possible pseudogenes because of the presence of insertions or deletions within the coding region. A phylogenetic analysis of the multiply aligned sequences (Fig. 4) demonstrates the varied nature of retroviral sequence elements within the human genome. An additional four clones contained artifactual sequences not related to reverse transcriptases. Three of the 27 clones contained a sequence identical to that of AMV reverse transcriptase, the enzyme used for cDNA synthesis, indicating the likely presence of DNA contamination in the enzyme preparation. In summary, our results demonstrate that hybrid primers can be used to isolate diverse members of multi-gene families simultaneously.

Our results can be compared with those obtained in two previous studies using the LPQG and YMDD reverse transcriptase regions

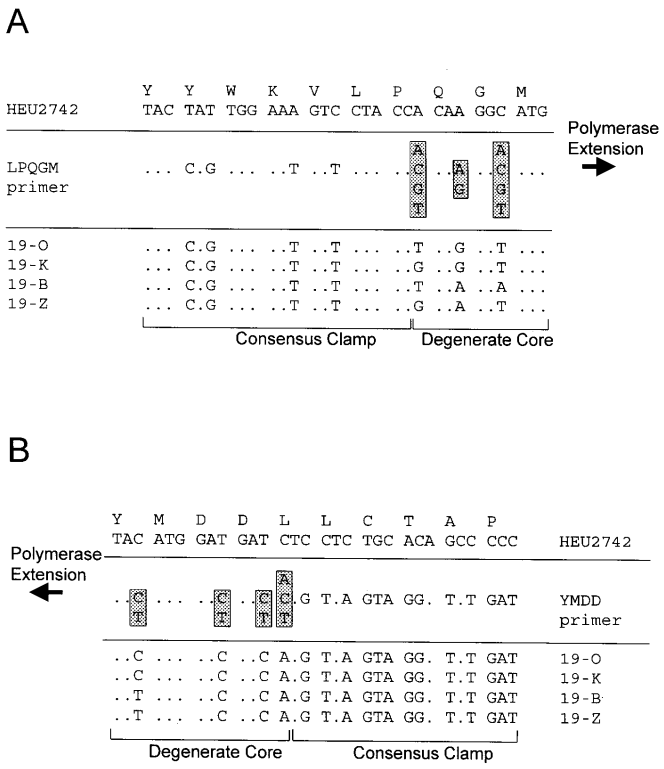

Figure 5. Analysis of hybrid primer utilization. The sequences of the hybrid primers, LPQGM and YMDD, incorporated into PCR products during the final amplification reaction of the experiment described in Figure 4, were determined from clones 19-O, -K, -B and -Z which contain a fragment of the retroviral element HEU2742 (GenBank #U27242). Nucleotide and amino acid sequences of the LPQGM (**A**) and YMDD (**B**) primer binding sites in HEU2742 are shown. The sequences of hybrid primer pools are aligned with the HEU2742 sequences and the sequences of degenerate codons in the primer pools are in shaded boxes. The direction of polymerase extension is indicated and the downstream YMDD primer is shown as its complement for clarity. Sequences from the incorporated primer for each clone are aligned with that of HEU2742, where identical residues are indicated (.).

for conventional degenerate primer design (2,18). In both studies, gel purification of PCR products was necessary. Nevertheless, in one study, only three of 17 clones were correct (2). In the other study, successful amplification was only obtained using purified viral template (18). In contrast, application of our hybrid primer method to minute amounts of genomic DNA present in formalin-fixed paraffin block sections yielded 48/52 correct clones from unpurified PCR products.

Analysis of hybrid primer utilization

To determine the utilization of hybrid primers during PCR amplification, we analyzed the sequences across the primers incorporated into four of the clones obtained with the LPQGM and YMDD primers. These four clones (19-B, -K, -O, -Z) corresponded to the human retroviral element HEU2742 whose sequence was available in GenBank. The sequences across the LPQGM and YMDD primer binding sites in HEU2742 were compared with the sequences obtained from the primers incorporated into the four different clones (Fig. 5A and B). In the core regions, the unknown template was found to encode the same invariant amino acid residues present in the alignment used to predict the primer. Consistent with the premise that multiple hybrid primers would participate in amplifying the correct target, six of the eight clone ends had incorporated primers with different sequences.

As expected, the sequences corresponding to the 5′ consensus region of the cloned primers were identical to one another but differed from the sequence of the HEU2742 template. In the case of the LPQGM primers, the 5′ consensus region matched the HEU2742 template sequence at 16/20 nucleotide residues. However, in the case of the YMDD primers, only 4/17 nucleotide residues in the consensus region matched the template. This poorly-matched 5′ clamp appears to have stabilized the 3′ core during the 55°C annealing step, because even a perfectly-matched core should have melted at 34°C (12).

Using the CODEHOP prediction program to isolate gene homologs from different genomes

Degenerate PCR primers have been used with limited success for obtaining eukaryotic C^5 DNA methyltransferases. For example, the mouse DNA methyltransferase was used to design degenerate PCR primer pools that led to isolation of the *Arabidopsis thaliana MET1* gene based on typical low stringency amplification and purification of a gel fragment of the correct size (19). These primers were used in an attempt to obtain DNA methyltransferases from other plants, including oak, salal and rhododendron; however, no bands of the correct size (except for *Arabidopsis*) were resolved (data not shown). Therefore, we judged that eukaryotic C^5 DNA methyltransferases represent a challenging family for isolation of new members by PCR.

A program to design consensus-degenerate hybrid oligonucleotide primers (CODEHOP) was written that applies the general rules used to design primers in the previous sections. Program input is a set of blocks and output is a primer map that lists CODEHOPs which fulfill specified stringency criteria. To test the CODEHOP strategy on the higher eukaryotic C^5 DNA methyltransferases, all eight available sequences were presented to BlockMaker (17), resulting in a set of six blocks corresponding to the six well-known conserved regions of these proteins (7; 20). Two of the sequences are from the ‘chromomethylase’ subfamily of predicted proteins in *A.thaliana* and its closest relative, *Cardaminopsis arenosa* (21). The other six sequences comprise a set of presumed DNA methyltransferase orthologs from animals (sea urchins to humans) and a plant (*A.thaliana* MET1). To bias the primers towards chromomethylases, the two members of this subfamily were upweighted by an arbitrary factor of four times the sequence weights, which are automatically provided by BlockMaker to reduce redundancy of close relatives (8). Using the C^5 DNA methyltransferase blocks as input, three pairs of optimal primers were identified. Two pairs would potentially amplify a sufficiently short region in the known chromomethylase genomic sequences (<500 bp) to be of practical use. For one of the predicted primers, the primer design strategy is shown (Fig. 6).

One CODEHOP pair produced complicated patterns of bands in various plant samples and even in the presumed negative control from *Drosophila melanogaster*, so products were not analyzed in detail. The other CODEHOP pair amplified products of the expected size (~250 bp) using DNAs from *A.thaliana*, broccoli, rhododendron, salal, stonecrop, oak and barley. The PCR reaction product from each sample was used for cloning into a plasmid vector without purification. Sequence analysis revealed that correct amplification of a putative chromomethylase occurred for *A.thaliana* (2/2 clones), broccoli (2/2 clones) rhododendron

A

MTDM_CHICK	E	M	L	C	G	G	P	P	C	Q	G
MTDM_HUMAN	E	M	L	C	G	G	P	P	C	Q	G
MTDM_MOUSE	E	M	L	C	G	G	P	P	C	Q	G
MTDM_XENLA	E	M	L	C	G	G	P	P	C	Q	G
MTDM_PARLI	E	L	L	C	G	G	P	P	C	Q	G
MTDM_ARATH	D	F	I	N	G	G	P	P	C	Q	G
MTCH_CARAR	Y	S	V	C	G	G	P	P	C	Q	G
MTCH_ARATH	Y	T	V	C	G	G	P	P	C	Q	G

B

Most AA	Y	M	V	C	G	G	P	P	C	Q	G
Most Codon	T A **C**	**A T G**	**G T T**	**T G T**	**G G A**	**G G A**	**C C T**	C C T	T G T	C A A	G G A

C

A	0100 8	74 0 12	11 0 15	4 4 0	0 0 37	0 0 37	0 0 33	0 0 33	0 0 0	0100 53	0 0 37
C	0 0 44	2 40 17	14 0 22	0 0 43	0 0 13	0 0 13	100100 12	100100 12	0 0 43	100 0 0	0 0 13
G	23 0 8	0 6 49	69 0 25	0 96 0	100100 14	100100 14	0 0 17	0 0 17	0100 0	0 0 47	100100 14
T	77 0 39	24 54 23	7100 38	96 0 57	0 0 36	0 0 36	0 0 38	0 0 38	100 0 57	0 0 0	0 0 36
Most Degen	K A N	H B N	N T N	W R Y	G G N	G G N	C C N	**C C N**	**T G Y**	**C A R**	**G G** N
Degen	2 1 4	2 3 4	4 1 4	2 2 2	1 1 4	1 1 4	1 1 4	1 1 4	1 1 2	1 1 2	1 1 4

Figure 6. The highlighted CODEHOP, consisting of an 11 residue 3′ degenerate core and a 19 residue 5′ consensus clamp, was predicted from the alignment shown. (**A**) Portion of a block alignment of eight sequences. MTCH_ARATH and MTCH_CARAR are chromomethylases from *A.thaliana* and *C.arenosa*, respectively; these were given weights four times those assigned by the position-based sequence weighting method (8) in order to bias the hybrid primers towards them. (**B**) The consensus residues from the amino acid PSSM for the block (which is not shown), and the corresponding most common codons according to the codon usage table for *A.thaliana*. (**C**) DNA PSSM with the most degenerate residue and degeneracy value at each position. The best suggested CODEHOP has degeneracy of 16 in the core region and the degenerate residues are underlined; the clamp region is drawn from the most common codons in (B), also underlined.

(2/2 clones), salal (1/1 clones), stonecrop (2/2 clones) and oak (1/2 clones) (Fig. 7A). A dendrogram of the translated sequences shows that the branch lengths of the putative chromomethylases from these dicot plants are almost two-fold longer than the branch lengths of animal C^5 DNA methyltransferases, ranging from mammals to sea urchins (Fig. 7B). Therefore, this CODEHOP pair successfully amplified chromomethylases that appear to be more diverse than the orthologous set of DNA methyltransferases from vertebrates and echinoderms.

Interestingly, the two broccoli clones came from different chromomethylase-like genomic sequences. The dendrogram indicates that one broccoli sequence is more closely similar to the CMT1 sequences of other mustards, *A.thaliana* and *C.arenosa,* than it is to the other plants, as expected. However, the other broccoli sequence, designated CMT2, groups with the other plants. This result was confirmed by using a broccoli CMT2 CODEHOP-based clone to select by filter hybridization an *A.thaliana* genomic clone containing a CMT2 homolog. Sequencing revealed that *A.thaliana CMT2* has an almost identical exon/intron structure to *CMT1* and encodes a chromomethylase that aligns with 43% amino acid identity over the full length of CMT1, with a CMT2-specific N-terminal extension (L.Comai, C.M.McCallum and S.Henikoff, unpublished results).

One of three clones from barley (a monocot with a 5000 Mb genome) yielded a sequence that is significantly different from *A.thaliana* MET1 but not from the known animal DNA methyltransferases, which are thought to be orthologous to MET1. The presence of this sequence in the crude barley DNA preparation was confirmed by subsequent amplifications using specific primers internal to the CODEHOP pair. However, these internal primers failed to amplify any specific product from a highly purified barley DNA preparation derived from a different source (data not shown). It therefore appears that the non-plant-like sequence arose from contamination of our first barley sample with an organism unrelated to barley, such as a fungus growing on the barley. Regardless of the source of this sequence, it is interesting that a member of the orthologous set of eukaryotic C^5 DNA methyltransferases was identified using primers biased towards chromomethylases, indicating that CODEHOPs are able to amplify DNA methyltransferases from two diverged subfamilies in a background of complex genomic DNA.

DISCUSSION

Isolation of an unknown sequence related to known sequences is a powerful method for investigating biological function. The sequence of an unknown protein in one organism may be homologous to those of known proteins from different organisms, or may be related to a known protein sequence belonging to a multigene family within an organism. In many cases, low-stringency hybridization or PCR methods have succeeded in obtaining such desired genes. However, as the degree of protein similarity decreases, so does success in gene isolation. When only a single sequence is known, low-stringency hybridization is used, although a fairly long region of similarity may be needed. Moreover, considerable effort is required to determine whether a candidate clone is a correct one. If a family of proteins is available, then consensus or degenerate PCR methods may be used, because regions of high sequence similarity can be identified and utilized in the design of PCR primers. PCR methods are not only faster and easier than low-stringency hybridization, but product size and homogeneity can also be used to judge probable success. However, consensus primers may be too dissimilar to an unknown target to efficiently anneal to the original template, and degenerate primers may be too dissimilar to each other to efficiently amplify the synthesized product. In either case, mismatches in oligonucleotide annealing are typically limiting; however, ignorance of how mismatches affect annealing (22) has resulted in primer designs that are largely subjective and that must be optimized by time-consuming trial-and-error testing.

A

```
                Oak  SGFNRFRNKENPLDDEKNKQLIVYMNLVQHLKPRYVLMENVVDLVKFANGFLGRYALGRLIG
          Stonecrop  SGFNRFRNKENPLEDTKNEQIIVYMDIVNYLKPRFVLMENVVDILKLSQGFLGRYALARLIQ
        Rhododendron  SGFNRFRNVTEPLSDPKNHQLEVFMDIVEFLKPRFVLMENVVDIFKLAGGVLGCYAVARLVS
               Salal  SGFNRFRNVTEPLKDPKNQQLIVFMDIVEFLKPRFVLMENVVDIFKLAGGVLGCYAIARLVS
      Broccoli CMT2  SGYNRFRDVDSPLTDERNQQIVVFMDIVEYLKPKFVLMENVVDILRLDKGSLGRYALSRLVD
      Broccoli CMT1  SGFNRFRNNKAPLEDEKNQQLLVYMNIIDYLKPNYVLMENVVDLLKFSKGFCARYAVARLVA
   A. thaliana CMT1  SGRNRYRNNEAPLEDQKNQQLLVFLDIIDFLKPNYVLMENVVDLLRFSKGFLARHAVASFVA
    C. arenosa CMT1  SGYNRFRNKQAPLEDKKNQQLLVFLDIIDFLKPSYVLMENVVDLLRFSKGYLARHAVASFVA
   A. thaliana MET1  SGMNRF  NQSSWSKVQCEMILAFLSFADYFRPRYFLLENVRTFVSFNKGQTFQLTLASLLE
         Barley (?)  SGMNRF  NSGQYSLFKNSLIVSFLSYIDFYRPKYFVMENVRNFVSFKRSMVLKLTLRCITR
        Sea Urchin   SGMNRF  NSREYSKFKNSLISSYLSYCDYYRPRFFLLENVRNFVSYKKNMVLKLALRCLIR
           Xenopus   SGMNRF  NSRTYSKFKNSLVVSYLSYCDYYRPKYFLLENVRNFVSFKKSMVLKLTLRCLVR
             Mouse   SGMNRF  NSRTYSKFKNSLVVSFLSYCDYYRPRFFLLENVRNFVSYRRSMVLKLTLRCLIR
           Chicken   SGMNRF  NSRTYSKFKNSLVVSFLSYCDYYRPRFFLLENVRNFVSFKRSMVLKLTLRCLVR
             Human   SGMNRF  NSRTYSKFKNSLVVSFLSYCDYYRPRFFLLENVRNFVSFKRSMVLKLTLRCLVR
```

B

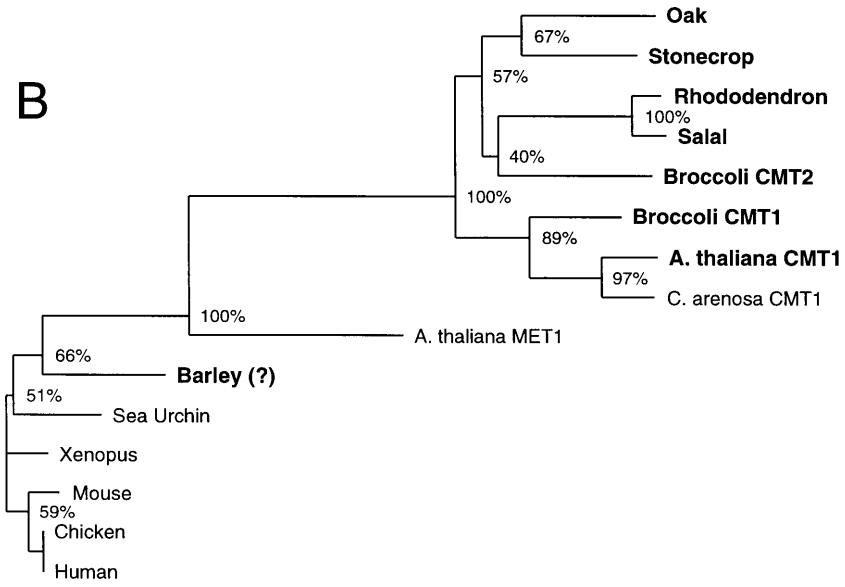

Figure 7. Alignment of higher eukaryotic DNA methyltransferases and translated PCR products (in bold) obtained using CODEHOP-designed primers (**A**) and the corresponding dendrogram showing bootstrap resampling percentages (**B**). GenBank accession numbers for amplified DNA sequences are AF47322–AF47328. PCR reactions were performed using primers designed by the CODEHOP program with BlockMaker MOTIF-generated blocks from the eight protein sequences listed in Figure 6 as input. The upstream primer was 5′-CATGGTTTGTGGAGGACCTCCNTGYCARGG-3′ (Fig. 6) and the downstream primer was 5′-TTGCATCATTC-CGAATCTACAYTGRTANYYCAT-3′. A hot-start was obtained by using Ampli-Taq Gold (Perkin-Elmer, 2.5 U/50 µl) and buffer with 4 mM $MgCl_2$ (Perkin-Elmer) with a 9 min pre-heating step at 94°C, followed by 40 cycles (30 s at 94°C, 30 s at 53°C and 30 s at 72°C) and a final 7 min, 72°C incubation.

Our novel hybrid strategy overcomes drawbacks of both consensus and degenerate methods by basing primer design on precisely-matched regions only. We presume that correctly amplified products are initially produced by precise matching of primer to template in the 3′ core and later by precise matching of primer to product in the 5′ clamp. The CODEHOP algorithm is aimed at minimizing mismatches between the consensus clamp and unknown templates, so that mismatches are unlikely to limit the application of our strategy to challenging problems. It seems more likely that our method is limited by the degeneracy of the 3′ core, which our algorithm optimally selects.

The practical utility of the hybrid method is demonstrated by successful amplification of unknown sequences that are too diverged from known sequences to be readily isolated by standard methods. In addition, the hybrid method was successful in amplifying unknown target sequences from sources containing small quantities of degraded nucleic acids, even single viral sequences present in a small minority of cells. In all cases, single PCR products of the correct size were observed by analytical agarose gel electrophoresis, so no gel purification was necessary. Our method was also successful in isolating diverse related products in a single reaction.

Although we rely on stabilization of the 3′ core by the presumably mismatched 5′ clamp in annealing to template, our data indicate that even poorly-matched clamps can be effective. This suggests that the actual sequence of the clamp is not always important, in which case annealing to template would be stabilized by any 5′ extension. It may be that the common practice of adding an arbitrary 5′ extension to a degenerate primer in order to introduce a restriction site is inadvertently responsible for many successful amplifications of unknown sequences in the past. Furthermore, the evident effectiveness of a clamp that is mismatched to template suggests that our hybrid strategy can be used for gene isolation when only short peptide sequences are

available for primer design. In such cases, the 3′ core would correspond to reverse translation of the least degenerate 3–4 amino acid region, and the 5′ clamp could extend beyond available sequence with arbitrarily chosen residues.

As sequence databanks grow and more sequences are classified into known families, the conserved protein regions become better delineated; this can aid in PCR primer design. At present, the Blocks Database (v. 9.3) contains 3417 alignment blocks representing 932 protein families, with an average of 23 sequences per family (16). Blocks from relatively similar sequences have been previously used for designing effective degenerate PCR primers (5). However, for more diverged families, there are too few consecutive invariant and highly conserved residues with low codon degeneracy to design efficient degenerate or consensus PCR primers. Because our hybrid strategy requires no more than four consecutive highly conserved amino acids, it can be more generally applied to these diverse protein families.

We have implemented the CODEHOP method as a computer program that is available for interactive use on the WWW. Previous programs have been introduced to design PCR primers to match known templates (11,23–25). When designing primers to unknown templates, other programs have been developed to minimize potential mismatches by identifying regions of low variability and codon degeneracy (26). Unfortunately, no theory or systematic method exists to guide primer design for unknown templates (22). Our new strategy, however, provides guidelines for design of efficient primers by limiting the degeneracy to just the 3′ 11–12 nucleotides of a primer and stabilizing annealing with a long consensus clamp. Moreover, the CODEHOP program utilizes all of the information available in the input alignment and takes into account the codon usage of the target genome to aid in primer design. The program first converts protein multiple sequence alignments into scoring matrices that consider sequence redundancy and amino acid conservation. These matrices are then converted to DNA frequency matrices tailored by organism-specific codon usage tables, and these DNA matrices are searched for optimal hybrid primers. Primers are displayed on a map that shows the level of degeneracy of the 3′ core and the maximum annealing temperature of the 5′ clamp, the length of which is based on the nearest-neighbor free energy method (12).

WWW implementation of the CODEHOP program has allowed it to be directly linked to the BlockMaker site for producing suitable multiple alignments from related protein sequences submitted by the user. The program is used interactively, so that parameters may be varied if needed: users can adjust the desired annealing temperature, the degree of degeneracy and the cut-off frequency level for bases allowed in the 3′ core region. Because there are no mismatches between primers and PCR products in the 5′ clamp region, stringent annealing conditions may be used, thus minimizing mispriming. We have found that annealing temperatures as high as 65°C can yield correct product, although stepwise reduction of the annealing temperature down to 50°C may lead to successful amplification without unacceptable background if no product is detected initially. A useful feature of the program is the ability to manually modify alignments or weights as desired. For example, reweighting sequences in order to favor certain ones was employed in designing CODEHOP pairs for the preferential amplification of plant chromomethylases relative to other C^5 DNA methyltransferases.

We have found that the CODEHOP method can be extended to even more divergent target sequences by using higher degeneracies and purifying PCR products of the anticipated size on high resolution polyacrylamide electrophoretic gels (T.M.R., unpublished results). We are currently testing the use of touchdown PCR (27) and polymerase time-release with the CODEHOP method (S.H., unpublished data). Other possible enhancements might increase the effectiveness of our method, such as changes in the program that would vary the length of the degenerate core or score the consensus clamp. These and other refinements should lead to even more efficient isolation of distantly-related unknown sequences than can be obtained at present.

ACKNOWLEDGEMENTS

This work was supported in part by a grant to T.M.R. from the M.J.Murdock Charitable Trust and by a grant to S.H. from NIH. S.P. is a Howard Hughes Medical Institute Fellow of the Life Sciences Research Foundation.

REFERENCES

1 Burglin,T.R., Finney,M., Coulson,A. and Ruvkun,G. (1989) *Nature*, **341**, 239–243.
2 Wichman,H.A. and Van Den Bussche,R.A. (1992) *Biotechniques*, **13**, 258–265.
3 Kim,Y.-J. and Baker,B.S. (1993) *Mol. Cell. Biol.*, **13**, 174–183.
4 Robertson,H.M. (1993) *Nature*, **362**, 241–245.
5 D'Esposito,M., Pilia,G. and Schlessinger,D. (1994) *Hum. Mol. Genet.*, **3**, 735–740.
6 Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) *Nucleic Acids Res.*, **22**, 4673–4680.
7 Posfai,J., Bhagwat,A.S., Posfai,G. and Roberts,R.J. (1989) *Nucleic Acids Res.*, **17**, 2421–2435.
8 Henikoff,S. and Henikoff,J.G. (1994) *J. Mol. Biol.*, **243**, 574–578.
9 Henikoff,J.G. and Henikoff,S. (1996) *Comput. Appl. Biosci.*, **12**, 135–143.
10 Nakamura,Y., Gojobori,T. and Ikemura,T. (1997) *Nucleic Acids Res.*, **25**, 244–245.
11 Rychlik,W. and Rhoads,R.E. (1989) *Nucleic Acids Res.*, **17**, 8534–8551.
12 Rychlik,W., Spencer,W.J. and Rhoads,R.E. (1990) *Nucleic Acids Res.*, **18**, 6409–6412.
13 Rose,T.M., Strand,K.B., Schultz,E.R., Schaefer,G., Rankin,G.W.J., Thouless,M.E., Tsai,C.C. and Bosch,M.L. (1997) *J. Virol.*, **71**, 4138–4144.
14 Sambrook,J., Fritsch,E.F. and Maniatis,T. (1989) *Molecular Cloning: A Laboratory Manual*, 2nd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
15 Ausubel,F.M., Brent,R., Kingston,R.E., Moore,D.D., Seidman,J.G., Smith,J.A. and Struhl,K. (1994) *Current Protocols in Molecular Biology.* Wiley, New York.
16 Lisitsyn,N., Lisitsyn,N. and Wigler,M. (1993) *Science*, **259**, 946–951.
17 Henikoff,S., Henikoff,J.G., Alford,W.J. and Pietrokovski,S. (1995) *Gene*, **163**, GC17–GC26.
18 Donehower,L.A., Bohannon,R.C., Ford,R.J. and Gibbs,R.A. (1990) *J. Virol. Meth.*, **28**, 33–46.
19 Finnegan,E.J. and Dennis,E.S. (1993) *Nucleic Acids Res.*, **21**, 2383–2388.
20 Cheng,X., Kumar,S., Posfai,J., Pflugrath,J.W. and Roberts,R.J. (1993) *Cell*, **74**, 299–307.
21 Henikoff,S. and Comai,L. (1998) *Genetics*, In press.
22 Rubin,E. and Levy,A.A. (1996) *Nucleic Acids Res.*, **24**, 3538–3545.
23 Lowe,T., Sharefkin,J., Yang,S.Q. and Dieffenbach,C.W. (1990) *Nucleic Acids Res.*, **18**, 1757–1761.
24 Hillier,L. and Green,P. (1991) *PCR Meth. Appl.*, **1**, 124–138.
25 Engels,W.R. (1993) *Trends Biochem. Sci.*, **18**, 448–450.
26 Dopazo,J., Rodriguez,A., Saiz,J.C. and Sobrino,F. (1993) *Comput. Appl. Biosci.*, **9**, 123–125.
27 Don,R.H., Cox,P.T., Wainwright,B.J., Baker,K. and Mattick,J.S. (1991) *Nucleic Acids Res.*, **19**, 4008.